In-context Learning of Linear Dynamical Systems with Transformers: Approximation Bounds and Depth Separation

*Frank Cole**, *Yuxuan Zhao**, Yulong Lu, Tianhao Zhang

School of Mathematics, University of Minnesota

NEURAL INFORMATION PROCESSING SYSTEMS

Results:

Theorem 1: There exists a linear transformer with $\mathcal{O}(\log(T))$ layers that satisfies:

$$L_T(\theta) = \mathcal{O}\left(\frac{\log(T)}{T}\right)$$

Interpretation: *Transformers with $\mathcal{O}(\log(T))$ depth can in-context learn linear dynamical systems, with approximation error matching that of the least-squares predictor.*

Theorem 2: Take $d = 1$. Then single-layer transformers cannot in-context learn linear dynamical systems: there is a $c > 0$ such that

$$\lim_{T\to\infty} \inf_{\theta \in \text{1-layer TF}} L_T(\theta) \geq c.$$

Interpretation: *A single linear attention layer is fundamentally limited in its ability to in-context learn linear dynamical systems, even with arbitrarily long context. This result uncovers a surprising distinction between in-context learning with IID and non-IID data.*

Future directions: Reduce number of layers required for upper bound, extend theory to nonlinear dynamical systems.

SCAN ME

Setup:

Model input: noisy LDS $(x_0, \ldots, x_T)$, where $x_T = Wx_{T-1} + \xi_T$, $\xi_T \sim \mathcal{N}(0, \sigma^2 I_d)$, $x_0 = \mathbf{0}_d$, and $W \sim \mathcal{P}_W$

Model output: estimate of the conditional mean of x_{T+1}:
$\hat{x}_{T+1} = \mathcal{T}_\theta(x_0, \ldots, x_T) \approx Wx_T$

Linear transformer layers: $Z \mapsto \hat{Z} = W_{\text{MLP}}\left(Z + \frac{1}{T} W_P Z Z^T W_Q Z\right)$

where $Z = \begin{pmatrix} 0 & \ldots & 0 \\ 0 & \ldots & 0 \\ x_1 & \ldots & x_T \\ x_0 & \ldots & x_{T-1} \end{pmatrix}$

Question: How well can linear transformers in-context learn linear dynamical systems, as measured by the worst-case loss function:

$$L_T(\theta) := \sup_{w_{\min} I_d \prec W \prec w_{\max} I_d} \mathbb{E}_\xi \left[\left\| \hat{x}_{T+1} - Wx_T \right\|^2 \right].$$

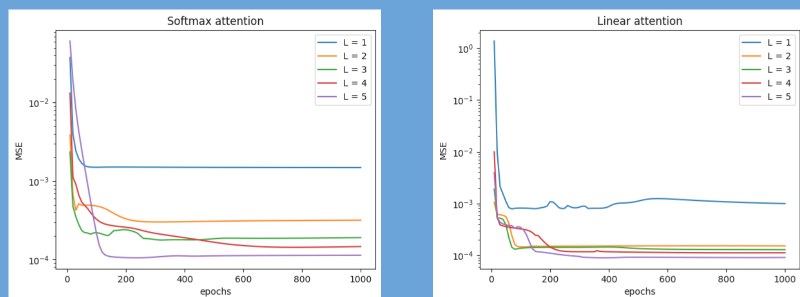

Figure: plotting test error of deep transformers, with and without softmax, as a function of the number of layers.

Proof techniques:

Theorem 1: we construct a transformer to implement the predictor $\hat{x}_{T+1} = \hat{W}_T x_T$, where $\hat{W}_T$ is the *least-squares matrix*:

$$\hat{W}_T := \left(\frac{1}{T}\sum_{i=0}^{T-1} x_{i+1}x_i^T\right)\left(\sum_{i=0}^{T-1} x_i x_i^T\right)^{-1}.$$

To approximate the inverse covariance, and drawing inspiration from [1], we construct a transformer to implement one step of the *modified Richardson iteration* $z_{\ell+1} = \left(A - \alpha I_d\right) z_\ell + \alpha b$, which is used to solve the linear equation $Az = b$. To prove our approximation bound, we combine the fast convergence of the Richardson iteration with statistical guarantees of the least squares matrix [2].

Theorem 2: To prove the lower bound in 1D, we directly compute the limit of the individual loss function $\ell(\theta, w) := \lim_{T\to\infty} \mathbb{E}\left(\hat{x}_{T+1} - wx_T\right)^2$. We then prove that $\inf_\theta \sup_w \ell(\theta, w) > c$ for some $c > 0$. Finally, we use the regularity of the loss to exchange limits and supreme.

[1] J. Von Oswald, E. Niklasson, E. Randazzo, J. Sacramento, A. Mordvintsev, A. Zhmoginov, and M. Vladymyrov. Transformers learn in-context by gradient descent. In International Conference on Machine Learning, pages 35151–35174. PMLR, 2023a

[2] N. Matni and S. Tu. A tutorial on concentration bounds for system identification. In 2019 IEEE 58th conference on decision and control (CDC), pages 3741–3749. IEEE, 2019.